You Can Judge an Artist by an Album Cover: Using Images for Music Annotation

Jānis Lībeks
Swarthmore College
Swarthmore, PA 19081
Email: jlibeks1@cs.swarthmore.edu

Douglas Turnbull
Department of Computer Science
Ithaca College
Ithaca, NY 14850
Email: dturnbull@ithaca.edu

***Abstract*—While the perception of music tends to focus on our acoustic listening experience, the *image* of an artist can play a role in how we categorize (and thus judge) the artistic work. Based on a user study, we show that both album cover artwork and promotional photographs encode valuable information that helps place an artist into a musical context. We also describe a simple computer vision system that can predict music genre tags based on content-based image analysis. This suggests that we can automatically learn some notion of artist similarity based on visual appearance alone. Such visual information may be helpful for improving music discovery in terms of the quality of recommendations, the efficiency of the search process, and the aesthetics of the multimedia experience.**

I. Introduction

Imagine that you are at a large summer music festival. You walk over to one of the side stages and observe a band which is about to begin their sound check. Each member of the band has long unkempt hair and is wearing black t-shirts, black boots, and tight black jeans with studded black leather belt. It would be reasonable to expect that they play heavy metal. Furthermore, it would be natural for you to check out a different stage before you hear them play a single note if you are not a fan of heavy metal music. This suggests that the outward appearance, or *image*, of an artist can play a role in how their music is received by audiences. Whether this image is carefully constructed by a public relations consultant or results from unintentional lifestyle habits of the performer, it encodes valuable information that helps place the artist into a musical context.

To this end, we are interested in exploring the relationship between music and music-related images (e.g., album cover artwork & promotional photographs of artists - see figure 1) for a number of reasons. Currently, popular music discovery engines, such as Pandora[1] and Last.fm[2], rely on such images to make their web and mobile applications visually appealing. That is, streaming both music and associated images provide a listener with an engaging multimedia experience.

While improving aesthetics is important for music discovery [7], our work focuses on using techniques from computer vision to make additional use of music-related images. First, we propose a new measure of music similarity based on *visual* appearance. Such a measure is useful, for example, because it allows us to develop a novel music retrieval paradigm in which a user can discover new artists by specifying a query image. Second, images of artists also represent an unexplored source of music information that is useful for the automatic annotation of music: associating semantic *tags* with artists [17]. Once annotated, an artist can be retrieved using a text-based query much like web pages are retrieved when using a typical Internet search engine (e.g., Yahoo!, Google). Finally, music-related images provide us with meaningful visual representation of sound. This is important when considering that it requires much less time to browse a large collection of images than to listen to a few short clips of music.

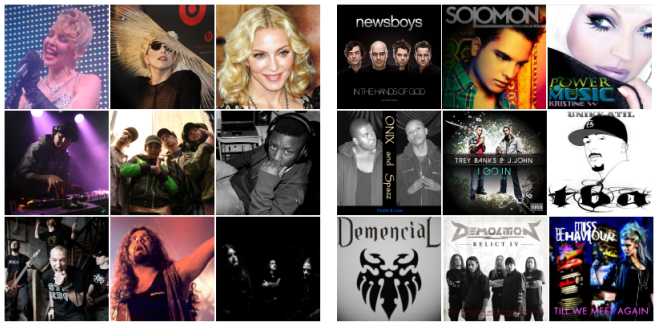

Fig. 1. Illustrative promotional photos (left) and album covers (right) of artists with the tags *pop* (1^{st} row), *hip hop* (2^{nd} row), *metal* (3^{rd} row). See Section VI for attribution.

In this paper, we describe an image annotation system that can both compute artist similarity and annotate artists with a set of genre tags based on album cover artwork or promotional photographs. Our system is based on a recently-proposed baseline approach called Joint Equal Contribution (JEC) [12]. JEC incorporates multiple forms of low-level color and texture information and has been shown to outperform numerous state-of-the-art approaches on standard benchmark image data sets. In order to use this approach of artist annotation, we modify it in a straight-forward manner so that we can use multiple images per artist to significantly improve performance.

II. Related work

Over the last decade, there has been a growing interest in developing techniques for both computing music similarity and for annotating music with tags [5]. This body of work focuses on content-based audio analysis, as well as using other sources

[1]http://www.pandora.com
[2]http://last.fm

of music information, such as social tags, music scores, lyrics, web documents, and preference data (e.g. [18], [3], [10]). To the best of our knowledge, music-related images, such as album covers and promotional photos, have not been used for these tasks. However, computer vision has been employed for other tasks such as optical music recognition [1], identifying documents with music notation [2], and identifying lyrics in scores [4]. In addition, standard computer vision techniques have been applied to 2-D representations (e.g., spectrograms) of audio content for music identification and fingerprinting [9].

Within the extensive computer vision literature, there are two general tasks that are related to our work. First, content-based image retrieval (CBIR) involves computing similarity between pairs of images. Deselaers et al. [6] provide a recent survey of CBIR research and describe a number of useful image features, many of which are used in this paper. The second relevant task is image annotation. For this task, the goal is to annotate an image with a set of tags (e.g., "sky", "polar bear", "forest"). Makadia et al. [12] recently proposed a system that combines color and texture features using Joint Equal Contribution (JEC) as a baseline approach for this task. However, the authors unexpectedly found that this approach performs better than a number of (more complex) systems.We use JEC as the core of our artist annotation system but extend it to use multiple images of each artist.

IMAGE SIMILARITY

To compute image similarity between two images using JEC, we first compute seven separate distances between each pair of images.

A. Image Features

The first three distances are related to color information. For each image, we compute one color histogram over each of three color spaces: red-green-blue (RGB), hue-saturation-value (HSV), and LAB. The three color histograms are 3-dimensional histograms extracted on 16 equally spaced bins for each color channel. The interval of the bins is determined from the possible range of values for each channel in each of the respective color space. Each color histogram is represented as a $16^3 = 4096$ dimensional feature vector where each element of the vector represents the (normalized) count of pixels that fall into a color bin. As in Makadia et al., we calculate the L_1-distance when comparing two RGB or two HSV histograms, and calculate the KL-divergence when comparing two LAB histograms.

The other four distances are related to two types of texture information: Gabor and Haar features. For the Gabor features, a grayscale version of the image is convolved with complex Gabor wavelets at three scales and four orientations to create 12 response images. A histogram of response magnitudes is performed using 16 equally-spaced bins with experimentally-determined maxima values. Finally, the 12 histograms are concatenated, creating a final 192-dimensional feature vector. This representation is referred to as *Gabor* in this paper. A second Gabor feature, called *GaborQ*, is calculated by averaging the response angles of the 126x126 image over non-overlapping blocks of size 14x14, quantizing to 8 values and concatenating the rows of each of the twelve resulting 9x9 images, resulting in a 972-dimensional feature vector. We compute both Gabor and GaborQ distances for each pair of images by calculating L_1-distance.

For the Haar features, we take the three Haar filters at three different scales, and convolve them with a (downsampled) 16x16 pixel grayscale version of the image. The simple concatenation of the response image, a 2304-dimensional vector, was called *Haar*. A second quantized version, referred to as *HaarQ* is found by changing each image response value to 1, 0 or -1 if the initial response value is positive, zero, or negative, respectively, again, producing a 2304-dimensional vector. Again, we calculate the Haar and HaarQ distance by computing the L_1-distance between pairs of Haar and pairs of HaarQ vectors.

	Genre Tags				
Artists closest to Daft Punk	electronic	dance	pop	house	classical
Astral Projection	1	0	0	0	0
Deadmau5	1	1	0	1	0
Eurythmics	1	1	1	0	0
Einstürzende Neubauten	1	0	0	0	0
Tags Predicted using JEC	1.0	0.5	0.25	0.25	0.0
Real Tags for Daft Punk	1	1	0	1	0

TABLE I
IMAGE-BASED ARTIST ANNOTATION USING TAG PROPAGATION. FOR A GIVEN SEED ARTIST (E.G. DAFT PUNK), WE RETRIEVE THE GROUND TRUTH TAG ANNOTATION VECTORS FOR THE ARTISTS WITH THE MOST SIMILAR IMAGES (E.G., ASTRAL PROJECTION, DEADMAU5, ...) AND THEN AVERAGE THEIR ANNOTATION VECTORS TO CALCULATE A PREDICTED ANNOTATION VECTOR.

B. Joint Equal Contribution

To combine the distances of the seven features using JEC, we normalize the distances for a feature by the maximum distance between any pair of images. This results in image-image distances in the range [0,1], where 0 denotes that the two images are the same, 1 denotes the most dissimilar pair of images. To combine feature vectors, we average the seven normalized distances over each pair of images. Note that JEC is a relatively simple approach because it does not require us to *learn* the parameters of (often complex) parametric models as is common in other approaches to image annotation. This makes the algorithm relatively easy to implement and fast to compute.

III. ARTIST ANNOTATION EXPERIMENTS

In this section, we explore image-based music similarity by considering the problem of annotating artists with genre tags. That is, we assume that two artists are similar if they are associated with a similar set of genres. In musicological terms, a *genre* encodes both auditory and cultural similarities betweens artists [15]. In information retrieval research, Lamere

and Celma find in a user study that computing artist similarity based on social tags lead to better music recommendations than when determined by human experts or audio content analysis [11]. They point out the majority of their social tags are music genres. McFee et al. also find that genre tags in particular are extremely useful for predicting artist similarity as determined by collaborative filtering [14].

Our genre tags are provided by Last.fm and are determined by a large number of individuals through a social tagging mechanism (i.e., "wisdom of the crowds.") Our system works by first finding visually similar artists to a given seed artist, and then *propagating* genre labels from these artists to the seed artist [10]. We argue that if the true genre tags for the seed artists are related to the propagated tags from the visually-similar artists, then our system is correctly finding some notion of music similarity based solely on visual appearance. We note that genre is a common surrogate for music similarity in information retrieval research (e.g., [15], [16].)

A. Data

Using Last.fm, we collect two separate image data sets (album covers & promotional photos) and a set of genre tags for a large number of artists. First, we create a vocabulary of genres by picking the 50 most popular genre tags on Last.fm. Due to the particular interests of the Last.fm community, some genres are rather broad ("classical", "country") where as others are somewhat specific ("melodic death metal", "trip-hop").

Next, for each tag we gather a list of the 50 most representative artists. We then collect a list of tags for each artist and retain the tags that appear in our vocabulary of 50 genre tags. This resulted in a set of 1710 unique artists and a binary tag matrix with an average of 4.74 tags per artist. Finally, for each artist, we attempt to download the 5 most popular promotional photos and 5 most popular album covers from Last.fm. Popularity is determined by Last.fm and appears to be related to the number of positive and negative votes that each image receives by their users. The downloaded images are pre-cropped to 126x126 pixels by Last.fm. This results in a set of 8417 album covers (average of 4.92 per artist) and 8527 promotional photos (average of 4.99 per artist). Finally, we clean up our data sets by removing duplicate artists. We also ignore duplicate images, which often appear, for example, when two or more artists in our data set appear on a compilation album or in a promotional photo together.

B. Tag Propagation

To evaluate the particular set of image similarity features, we compute a predicted tag vector for each artist. For each of the images associated with the artist, we find the 1-nearest neighbor image from the set of all other images in our data set. Next, we average the tag annotation vector for each of the matched artists. Thus, for each artist we have a predicted tag vector of values in the range [0,1], with 1 meaning that all neighboring artists are associated with the genre tag. See figure I for an illustrative example of the annotation process.

TABLE II
PERFORMANCE OF JEC IMAGE ANNOTATION ON ARTIST ANNOTATION TASK.

	Album Covers		Promotional Photos	
Feature	AUC	MAP	AUC	MAP
Random	*.500*	*.100*	*.500*	*.099*
RGB	.565	.132	.564	.139
HSV	.564	.131	.573	.151
LAB	.548	.127	.571	.140
Gabor	.547	.131	.527	.111
GaborQ	.571	.166	.517	.111
Haar	.578	.171	.544	.122
HaarQ	.580	.175	.524	.115
JEC	.598	.181	**.585**	**.159**
JEC without LAB	**.606**	**.192**	.581	.155

C. Evaluation

Next, we compute two information retrieval performance metrics for each tag: Mean area under the ROC curve (AUC) and mean average precision (MAP). For each tag, we start by ranking artists by their predicted tag value, and then calculate each performance metric using the ground truth tags for the artists. An ROC curve is a plot of the true positive rate as a function of the false positive rate as we move down this ranked list of artists. The area under the ROC curve is found by integrating the ROC curve. A perfect ranking (i.e., all the relevant artists at the top) results in an AUC equal to 1.0 and a random ranking produces and expected AUC of 0.5. Average precision (AP) is found by moving down our ranked list of artists and averaging the precisions at every point where we correctly identify a relevant artist. More details on these standard IR metrics can be found in Chapter 8 of [13]. To evaluate image similarity features, we compare the averages of the AUC and AP over all 50 genre tags.

When comparing image features, statistical significance is determined using a two-tailed paired t-test over the $n = 50$ tags with $\alpha = 0.05$. For example, we compare 50 differences in AUC scores for the 50 tags when comparing, say, randomly ranking songs verses ranking songs based RGB information. When one ranking system consistently outperforms another ranking system according to a t-test on these 50 differences, we say that that the system is significantly better.

D. Results

First, we explore our set of image features to determine which features are most appropriate for artist annotation. As in Makadia et al. [12], we consider the performance of each image similarity feature separately, and then calculate the performance when we combine features. The results for the seven image similarity features as well as the combined JEC approach are listed in table II.

We find that all of the features perform significantly better than random. It is interesting to note that, for the album cover data set, texture features (GaborQ, Haar, and HaarQ) perform best, whereas for the promotional photo data set, color features work best (RGB, HSV, LAB). More importantly, we note that the combination of all seven features using JEC

TABLE III
EFFECT OF USING MULTIPLE IMAGES OF EACH ARTIST WITH ALL SEVEN FEATURES (USING AUC).

# of images	1	2	3	4	5
Album Covers	0.518	0.547	0.570	**0.598**	0.570
Promotional Photos	0.517	0.542	0.567	**0.585**	0.566

performs significantly better than any individual feature. In addition, we explored removing individual features before computing JEC. In general this did not seem to significantly impact performance. However, for album covers, removing LAB significantly improved performance.

TABLE IV
AUC PERFORMANCE OF THE TEN BEST PERFORMING INDIVIDUAL TAGS USING THE JEC ON BOTH DATA SETS.

Album Covers	
Tag	AUC
Random	*0.50*
classical	0.74
metal	0.73
black metal	0.72
power metal	0.71
death metal	0.70
heavy metal	0.66
pop	0.65
dance	0.63
trip-hop	0.63
metalcore	0.63

Promotional Photos	
Tag	AUC
Random	*0.50*
melodic death metal	0.71
metal	0.69
power metal	0.68
death metal	0.67
metalcore	0.66
heavy metal	0.66
dance	0.64
classical	0.63
indie pop	0.63
thrash metal	0.63

In table III, we show that increasing the number of images used for each artist improves performance, but only up to the fourth image; adding the fifth most popular image of the artist decreases performance. Thus, we use the four most popular images for each artist for the rest of our analysis.

In table IV, we list the AUC performance of the ten best individual tags with JEC for both image data sets. Some of the tags, such as *metal*-related, *dance*, *pop*, and *classical*, tend to perform best (i.e., AUC > 0.63). It is also worth noting that six or seven of the ten most successful tags for both data sets contain the word *metal* (specifically, *metal*, *death metal*, *melodic death metal*, *thrash metal*, etc.), indicating that the top-level genre *metal* has a specific visual appearance that makes it easy to identify, based on our set of features.

On the other hand, we find that the performance of 6 individual tags is not statistically different from random when annotating artists using album covers[3]. When annotating based on promotional photos, we find that there are 10 tags that show no statistical difference from random. The common tags to both of these sets are *funk*, *country*, *reggae*, and *blues*.

IV. COMPARISON WITH HUMAN PERFORMANCE

In order to compare the performance of our computer vision system with human performance on the same task, we conducted a second experiment. Our study involved 397 English-speaking participants, each of whom was familiar with western popular music. Each individual participated in 12 trials which took a total of between five to eight minutes to complete. In each trial, we displayed either one or four images from either the album cover or promotional photo data set. We then asked the participant whether they recognized the artist based solely on these images, and then we asked them to select between one and five genre tags from our vocabulary of tags. We discarded the first two trials for each participant to account for the time necessary to become familiar with the format of the study. We then computed the average F-measure over all trials separated into the appropriate categories, as shown in table V. The F-measure is computed as the harmonic mean of precision and recall for each trial. In this case, precision is the number of tags that are both select by the user and found in the Last.fm tags divided by the number tags that are selected by the user. Recall is the number of tags selected by the user and found in the Last.fm tags divided by the total number of tags found from Last.fm.

[3]For each tag, we determine statistical significance ($\alpha = 0.05$) by comparing the AUC for the tag with $n = 1000$ bootstrapped estimates for AUC values based on different random rankings of the artists. This allows us to directly calculate a p-value from the empirical distribution of AUC values.

TABLE VI
THE FIVE BEST GENRE TAGS (BASED ON AVERAGE F-MEASURE) FOR HUMANS WHEN SHOWN 4 IMAGES OF AN ARTIST THAT WAS NOT RECOGNIZED.

Album Covers		
Tag	JEC	Human
Random	*0.09*	*0.09*
country	0.02	0.63
reggae	0.06	0.52
classical	0.28	0.49
soul	0.12	0.40
metal	0.49	0.39

Promotional Photos		
Tag	JEC	Human
Random	*0.09*	*0.09*
classical	0.21	0.59
rap	0.04	0.55
country	0.10	0.50
hip-hop	0.11	0.50
rnb	0.13	0.45

To compare with computer vision performance, we select the top five tags using JEC with tag propagation as described in section III-B based on either (one or four) promotional photos or album covers. Here, we compute statistical significance by performing a paired t-test on all trials ($\alpha = 0.05, n > 160$) performed by our system and humans for each of our four categories. As expected, the highest performance was attained when the participants recognized the artist in the images. More surprisingly, we see that our computer vision system (JEC) performs comparably with humans when they do not recognize the artist. For example, when considering four album covers, the participants' performance when they do not recognize the artist is significantly no different from the performance of our system. We observed that using four images instead of just one lead to a significantly better performance for both humans as well as our system. Note that both humans and our computer vision system perform well above random performance, thus signifying that information about musical genre *can* be extracted from music-related images.

Finally, we observe that humans perform well on some of the tags that our system performed worst on, such as *country*, *rap* and *reggae*, indicating that the respective artist images do in fact encode information about the musical style, however, our image annotation system does not adequately extract, encode or model this information. The five best tags based

TABLE V
RESULTS OF HUMAN STUDY ON THE TWO IMAGESETS. RESULTS FOR THE LAST FOUR COLUMNS ARE SHOWN AS F-MEASURES.

Images	Image Count	Artist Recognition Rate (%)	Human Recognized	Human Unrecognized	JEC	Random
Album Covers	1 Image	27.3	0.466	0.216	0.173	0.092
	4 Images	35.3	0.478	0.256	0.229	0.094
Promo Photos	1 Image	10.7	0.412	0.249	0.176	0.091
	4 Images	16.2	0.477	0.287	0.227	0.094

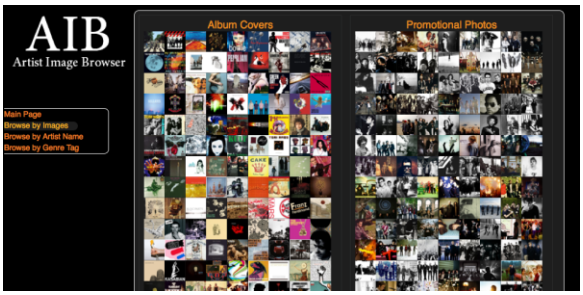

Fig. 2. Screenshot of the Artist Image Browser (AIB) website, illustrating the results of our experiment.

on human performance are listed in table VI.

V. DISCUSSION

In this paper, we have shown that we can automatically annotate artists with a large set of genre tags based solely on the analysis of album cover artwork and promotional photographs. We believe that this is an exciting new research direction for a number of reasons. First, it provides us with a novel *query-by-image* music discovery paradigm. To this end, we have developed a prototype web-based music image browser for exploring music similarity and annotation called the *Artist Image Browser*[4], see figure 2. We can also use this browser to visualize images that are the most representative of a genre tag based on content-based image analysis. Second, we have identified music-related images as a novel source of information for semantic music annotation [17]. Third, we have shown that images associated with music encode valuable information that is useful for contextualizing music. This has implication for how one might design a music discovery engine so as to provide a more effective and efficient means to *visualize* search results using relevant images.

It is important to note that the JEC system presented in this paper is only a *baseline* approach. That is, performance could be improved with a different selection of image features or an alternative image annotation model [8]. However, as we have shown in the previous section, the performance of our computer vision system is comparable to the performance of humans when they do not recognize the artist. This suggests that it might be hard to significantly improve our system without taking advantage of additional external information (e.g., audio content). Lastly, we recognize that our artist selection method biases our results towards more popular artists with a clearly established and commercialized *image*. Our system might perform worse on lesser known artists where outward appearance may be less carefully managed and manipulated.

[4]Artist Image Browser: http://jimi.ithaca.edu/aib/

VI. ACKNOWLEDGEMENTS

All images for Figure 1 are either from the Public Domain or are licensed under the Creative Commons Attribution 2.0 Generic license (attribution to David Rubin, Lori Tingey, David Shankbone, Briana (Breezy) Baldwin, Carlo Alberto Della Siega, Christian Borquez, Nathan Brown), or the Creative Commons Attribution 3.0 Unported license (attribution to Midwinter Music).

REFERENCES

[1] D. Bainbridge and T. Bell. The challenge of optical music recognition. *Computers and the Humanities*, 2001.
[2] D. Bainbridge and T. Bell. Identifying music documents in a collection of images. In *ISMIR*, pages 47–52, 2006.
[3] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman. A large-scale evalutation of acoustic and subjective music-similarity measures. *Computer Music Journal*, pages 63–76, 2004.
[4] J. A. Burgoyne, Y. Ouyang, T. Himmelman, J. Devaney, L. Pugin, and I. Fujinaga. Lyric extraction and recognition on digital images of early music sources. *ISMIR*, 2009.
[5] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.
[6] T. Deselaers, D. Keysers, and H. Ney. Features of image retrieval: An experimental comparison. *Information Retrieval*, 2008.
[7] J. Donaldson and P. Lamere. Using visualizations for music discovery. ISMIR Tutorial, October 2009.
[8] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models forimageauto-annotation. In *ICCV*, 2009.
[9] Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. In *CVPR*, June 2005.
[10] J.H. Kim, B. Tomasik, and D. Turnbull. Using artist similarity to propagate semantic information. *ISMIR*, 2009.
[11] P. Lamere and O. Celma. Music recommendation tutorial notes. ISMIR Tutorial, September 2007.
[12] A. Makadia, F. Pavlovic, and S. Kumar. A new baseline for image annotation. *ECCV*, 2008.
[13] C.D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
[14] B. McFee, L. Barrington, and G. Lanckriet. Learning similarity from collaborative filters. In *ISMIR*, 2010.
[15] C. McKay and I. Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved? *ISMIR*, 2006.
[16] E. Pampalk. *Computational Models of Music Similarity and their Application to Music Information Retrieval*. PhD thesis, Vienna University of Technology, 2006.
[17] D. Turnbull, L. Barrington, and G. Lanckriet. Five approaches to collecting tags for music. *ISMIR*, 2008.
[18] G. Tzanetakis and P. R. Cook. Musical genre classification of audio signals. *IEEE Transaction on Speech and Audio Processing*, 10(5):293–302, 7 2002.